The CMT Model of Free Will

08.01.2020

Louis Vervoort, Tomasz Blusiewicz

University of Tyumen, Russian Federation

Abstract: Here we propose a compatibilist theory of free will, in the tradition of naturalized philosophy, that attempts: 1) to provide a synthesis of a variety of well-known theories, capable of addressing problems of the latter; 2) to account for the fact that free will comes in degrees; 3) to interface with natural sciences, especially neurobiology. We argue that free will comes in degrees, as suggested by neuroscience. We suggest that a concept that can precisely 'measure' the variability of free will is that of 'theory', or rather, the capacity to make assumptions and use theories. Our model, then, highlights that free-willed actions are consciously monitored by the agent, through beliefs, assumptions and ultimately theories – hence the CMT model (for Conscious-through-Monitoring-through-Theories). The ultimate goal of this attempt at synthesis is to make the comparison with a variety of well-known accounts, and to show that our model can address several of their problems.

1. Introduction. Approach used.

The question whether humans have free will is an all-time classic of philosophy; a staggering number of 'great minds' from all disciplines have expressed their opinion on it. Two typical positions in the debate are hard determinism and libertarianism (cf. Fischer et al. 2007, Griffith 2013, Kane 2005 and 2012, Mele 2009, O'Connor and Franklin 2018, Pereboom 1997, Walter 2001, Watson 2003). Hard determinists reject free will on the hypothesis that all events, including our choices, are predetermined (fixed) by the laws of nature and the past; libertarians accept free will as a given and reject determinism. According to typical hard determinists free will would largely be an illusion, fuelled by our obvious *feeling* of being free, and entertained by our obvious incapacity to process all factors influencing us (and thus by our incapacity to predict in detail what will happen to us). Many professional philosophers however are compatibilists, maybe even a majority as claimed in (Pereboom 1997, p. 242): according to this position, determinism and free will are not in contradiction.

The model we elaborate here is compatibilist. We have presented arguments in favor of determinism elsewhere (Vervoort 2013), but our compatibilist view will actually appear to be independent of whether our universe is deterministic or indeterministic; hence our model could also appeal to proponents of indeterminism. Our theory is based on an advanced biophilosophical account of consciousness developed by Mahner and Bunge (1997). We adhere in the following to a naturalized philosophy, sympathetic to interaction with natural sciences (for recent works expounding the merits of naturalized philosophy, see e.g. Ladyman and Ross 2007, Mahner and Bunge 1997, Walter 2001). Thus we will in the conclusion briefly sketch how our model could be instrumental for some scientific disciplines, in particular neurobiology and computer science.

One goal is to construct a 'minimal' model in the sense that we aim to identify the minimal set of necessary and sufficient conditions to term an act 'free-willed' or 'free'. In this endeavor we use classic tools of analytic philosophy. However, we may be less typical in our attempt to synthesize prominent theories of free will. With synthesis, we mean the attempt (or its result) of showing how different established and efficient theories can be derived from one and the same unified model; more realistically, how the essential ingredients of different established models resort under one umbrella model. We believe that in the free will debate, millennia old, analytic philosophy has invested such enormous efforts and has made, especially the last fifty to sixty years, such progress in conceptual clarity, that efforts at synthesis are now well justified. This is so on the condition, of course, that the resulting synthetic model can solve worries where isolated accounts remain silent or powerless. Also, let us note that in natural science, especially physics, synthesis – or 'unification' as is it is termed there – of different assumptions and theories is recognized as the key principle for theory-building; we do not see any good reason why this could not be so in philosophy. Thus we insist that our model should be judged on its results, namely the fact that it can tackle problems that have been identified in competing modern theories, as shown in Section 7. Each of the ingredients or components of our model will appear necessary to address these problems.

Methodologically, our synthetic and 'naturalized' approach is inspired by, and can be compared to, the work of philosopher and neuroscientist Henrik Walter (2001), who proposes a component theory of maximal libertarian free will: a theory with essentially three requirements that a libertarian would demand of a 'free' act (in short, alternativism, intelligibility and agency). Walter has done an extensive

literature review to show that his component theory covers a wide spectrum of the most relevant modern theories and issues on free will; hence the relevance of benchmarking our model to Walter's[1].

Regarding the content of our model, our most conspicuous influences come from A. J. Ayer (1954/1997), and, especially, M. Mahner and M. Bunge (1997), as explained in Sections 4 and 5. These authors develop theories of which essential ideas can be traced back to such fore-runners of the philosophy of free will as Aristotle and Aquinas. In Section 6 we will show in detail how these models can be combined into our favorite model, which we propose to term the CMT-model of free will (for 'Conscious-through-Monitoring-through-Theories').

We will argue that one advantage of this model is that it captures a feature of free will we believe is essential, namely that free will comes in degrees. The variability of free will has been hinted to by neurobiologists (e.g. Brembs 2011), and convincingly analyzed by philosophers (O'Connor 2009); we will present our own arguments in Section 3, after explicitly stating our background assumptions in Section 2. Our model suggests, as far as we know for the first time, a qualitative 'measure' for this variability of free will (Section 6). As said, we will also show that this model can be related to, and synthetically explain salient properties of, very recent theories, notably those of Frankfurt (1996, 1988), Wolf (1990), Fischer and Ravizza (Fischer and Ravizza 1998, Fischer et al. 2007) and Walter (2001). Of course, we can here only make a succinct comparison with these elaborated theories (Section 7).

2. Background assumptions.

Maybe it is useful to explicitly state our background assumptions. First of all, we assume that our mental activity, our thoughts, choices, feelings etc., have a neurological, and ultimately chemical-physical basis in the brain: mental states correspond to neural (super)networks, mental acts are brain processes (Mahner and Bunge 1997, Walter 2001). Based on the famous Libet experiment and other neuroscientific experiments (showing for instance that 'free' choices can be predicted up to eight seconds before the subject's conscious action, cf. Soon et al. 2008), many, but certainly not all, neuroscientists favor a deterministic interpretation in the free will debate. This is an argument for us to look for a compatibilist

[1] Walter himself is not a libertarian; in his book he tests the three ingredients of libertarian free will against natural science, in particular neuroscience. He argues that neuroscience does not really support alternativism (the hypothesis that 'at the atomic scale', i.e. taking *all* physical facts into account, a human brain has, at any given time, real alternatives to choose between – cf. Walter 2001). So Walter is not in favor of strong libertarianism but shows that versions of intelligibility and agency do find support in neuroscience. We assume from the start, as compatibilists and determinists, that alternativism is eliminated.

model. In physics the orthodox position is that at the quantum level nature is probabilistic and therefore *in*deterministic. But a detailed scrutiny shows that this assumption is heavily metaphysically tainted and that the debate is undecided (Wuethrich 2011, Vervoort 2013). Simply stated, physical determinism is the assumption that given *all* physical data of the universe at any given moment t_0, and given the laws of nature, there is only one possible future after t_0. In other words, all events, including all mental events, are determined, caused, fixed by preceding events. Many determinists conclude from this we do not *really* have robust alternatives (alternativism is false) even if we feel we have. This would mean, in particular, that we do not have the 'ultimate' kind of free will that a typical libertarian assumes, the free will of a 'mind' that could make a real difference, could make a genuine choice at t_0, largely independently of the events preceding t_0. But some important form of free will may still exist; *which* form is the object of study of this article.

3. A preliminary observation: free will comes in degrees.

Technically speaking, our goal is to construct a definition of a 'free-willed' act by a human, in the form of necessary and sufficient conditions. Hence we aim at identifying the conditions C1, C2, C3,… such that 'Person P does act A of his own free will' if and only if (IFF) C1, C2, C3,… are satisfied. The Ci are all necessary conditions and they are jointly sufficient to imply free will. As announced, and remembering Occam's razor, we are looking for the *minimal*, most economic set {C1, C2, C3,…}. Finally, we look for conditions that are compatible with determinism. They will also appear to be compatible with *in*determinism; compatibilist models can have this feature. Therefore our definition would survive even if the universe would be indeterministic.

But we would also like our model to say something about a feature of free will that seems important to us, namely its variability. Consider your favorite dog, playing in the garden, and suddenly seeing the neighbour's cat appear. You shout: Aristotle, stay put ! For a fraction of a second Aristotle remains motionless, throws a guilty glance at you, and then hurries after the cat. Has Aristotle free will ? Could Aristotle have done otherwise, has he alternatives ? Maybe, but most philosophers agree that free will, and certainly the free will needed for moral responsibility[2], is essentially a human thing (if it exists at all).

[2] Note that this article is not about moral responsibility. We leave the link between free will and moral responsibility unspecified; we minimally declare we are interested in 'high level free will', typically understood by philosophers as required for moral responsibility.

But granted, one might say that Aristotle, the dog, possesses an embryonic version of free will. Has a new-born child free will ? A foetus ? No, most would say. When, then, does a child start having free will ? If a normal adult has free will, and if free will is linked to brain activity – a brain activity that grows continuously more complex when a foetus and new-born child evolve into an adult – than it seems one inevitably comes to the conclusion that free will comes in continuous degrees. This is, we believe, a rather innocuous conclusion, that has been noted before (O'Connor 2009) and with which many philosophers could agree, we believe – especially in the naturalized philosophy tradition[3]. Since this hypothesis will be highly instrumental in the following, we term it the 'degree-thesis'. Simply put: free will comes in degrees; in other words the mental capacity of free will is something that admits of (continuous) variations or gradations, something that a human (and generally speaking an animal) can have to greater or lesser extent – moreover an extent that can vary with time within an agent, as noted in (O'Connor 2009). Another way to understand this variability is to realize that animals have continuously evolved from bacteria over invertebrates, fishes, amphibians, reptiles, small mammals to primitive human species to modern Homo sapiens, and that somewhere in this evolution, paralleling a steady increase in brain complexity, free will must have appeared, not suddenly, but evolving in the same continuous way as the brain did. We will have much more to say on this in a moment. Indeed, we suggest that there exists a concept that can 'measure' this continuous progression.

4. A basic definition of free will (Ayer model).

Let us now start our main task and look for the simplest non-trivial compatibilist model of free will; this simplest model can however not solve certain problems and will have to be completed in a following stage. A considerable part of the academic debate on free will has turned around the 'principle of alternative possibilities' (PAP) (Frankfurt 1969, Kane 2005). According to the PAP, a person P does act A of his/her own free will only if P could have acted otherwise than doing A (this 'only-if' condition is then a necessary condition for free will). Most libertarians embrace the PAP as a proof that free will excludes determinism. But it is well known there are considerable problems with the characterization of free will provided by the PAP. Notably, Frankfurt has provided counterexamples, in which he presents

[3] O'Connor (2009) arrives at the conclusion of variability of free will by an independent route, notably by observing that agents can be, to lesser or greater extent, consciously aware of the factors guiding their choices. As will be seen further, this – in our view highly relevant – observation can well be fitted to our final model developed in Sections 6-7.

cases in which P could not have acted otherwise, while he still acts of free will; hence the PAP would be false (Frankfurt 1969, 1988). These arguments have been criticized in turn (e.g. Widerker and McKenna 2003); a detailed overview of the most recent developments in this intense debate can be found in (O'Connor and Franklin 2018). We believe it is fair to say that this matter is undecided.

A classic compatibilist attempt to defuse the worries related to the PAP was given in a well-known article by A. J. Ayer (1954/1997). The article famously argues that the capacity to choose as invoked in the PAP may well be guaranteed by the simple requirement that subject P acts *without constraint*. In its simplest expression, an act of free will, for Ayer, is an act that is not under constraint – an idea that traces back at least to Aristotle's Nicomachean Ethics (Book III). In some more detail, we can extract following analytic definition of free will from Ayer's article. 'P acts of his own free will' is essentially equivalent to 'P acts without constraint', more precisely:

DEF-A: 'P acts of his own free will' IFF

i. P is not compelled by other persons;
ii. there is no habitual ascendancy over P by other persons; and
iii. P acts voluntarily.

Ayer leaves the concept of 'voluntary act', again a notion that goes back to Aristotle (Nicomachean Ethics, Book III), rather imprecise; therefore we will have to come back to it in a moment. But from the text it follows that at least one ingredient of 'voluntary' is: not being under constraint of an abnormal mental condition such as a mania, a neurosis etc. (Ayer 1954 / 1997). Can an act that is not constrained and thus 'free' according to (i) – (iii) also be free in the sense of the PAP, in other words, could the unconstrained P also 'have done otherwise'? Ayer famously answers in the affirmative; he construes 'P could have acted otherwise' as follows: A) P was not constrained in the sense above (no-one compelled P and the act was done in a 'normal' psychological condition) and B) P *would* have acted otherwise *if P had so chosen*.

Sure, this construal is far from universally accepted (cf. the review in O'Connor and Franklin 2018, Section 2.2); but what is essential for us is that it appears that DEF-A *can be supplemented in order to address worries of much more sophisticated recent theories*, as argued in Section 7. Therefore we take DEF-A as a minimal analytic starting point. (Even if this is not essential at this point, one could then also adopt Ayer's construal of 'could…' as 'would… if…' as plausible, also because it is compatible with determinism (Ayer 1954/1997). Note that the same holds for clause A): P can be unconstrained *in the*

quite common sense (i) – (iii) even if everything is determined. In Ayer's words, free will should be contrasted with constraint, not determinism.) For the time being we retain: there are respectable arguments that an act of free will can be characterized in terms of absence of constraint in the sense (i) – (iii), with the proviso that we would like to know more about what 'voluntary' means in (iii).

Still, it is clear for other reasons that DEF-A cannot be the whole story. We can repeat an argument we already invoked above: 'real free will' should be something more than just 'unconstrained will that has the ability of choosing between options', because it seems that even animals could have this type of will – recall Aristotle the playful dog. But as most philosophers we ascribe 'real' free will only to humans.

What, then, is the feature that distinguishes humans most clearly from animals ? There is a long philosophical tradition related to this question, but it is presumably uncontroversial to identify consciousness, the capacity to think, rationality, as the key element of distinction. Then, something of this should enter the definition of free will. And indeed, *consciousness and/or rationality seem implicit in Ayer's concept of voluntariness*, as an ingredient of free will, as also follows from early analyses of the concept of voluntariness by Aristotle and Aquinas (cf. excerpts in the anthology Pereboom 1997, Ch. 1 and Ch. 5). We therefore should find a model for this essential but admittedly somewhat vague concept.

5. Voluntariness and consciousness (Mahner-Bunge model).

To analyze voluntariness, the most developed naturalized model we are aware of is proposed in Mahner and Bunge (1997). In their 'Foundations of Biophilosophy' the authors attempt a penetrating and wide-ranging approach with the aim to provide an axiomatic theory of key concepts of biology including psychobiology. Their approach and methods are eminently science-compatible, and their background assumptions coincide with ours (cf. Section 2 and below). In Chapter 6, on psychobiology, Mahner and Bunge (MB henceforth) propose an analysis of concepts as mind, mind-body interaction, consciousness, self, voluntariness, free will etc. We will give here a succinct overview of the notions we need for the derivation of our free will model. Clearly, MB's axiomatic and formalized approach is only one possible theoretical framework for the above concepts; but it seems to us their hypotheses and well-structured theory are programmatic and heuristically powerful, notably for comparison with neurobiology and computer science. We will only retain the essence, and occasionally make small changes as indicated.

MB closely link the mind with the neuronal states and networks in the human brain; mind and soul are not conceived of as immaterial, spiritual, perhaps immortal entities. In this theory the mind of an

animal is construed as the union, the *set* of all mental processes of its brain[4]. More precisely, Mahner and Bunge propose following definition, reproduced here literally (MB 1997 p. 205):

> **DEF-MB-1**. Let P denote a plastic neuronal supersystem of animal b of species K. Then the *mind* of b during the period τ is the union of all the mental processes (specific functions π_s) that components of P, i.e., plastic neuronal systems n, engage in during τ. More precisely,
>
> $$M(b, \tau) = \bigcup_n \pi_s(n, \tau).$$

The other key concepts, plastic neuronal (super)system, mental process, mental function, are also defined in (MB 1997 Ch. 6). However simple, this theory allows to address some of the old, paradigmatic problems of the philosophy of mind. One such question is: where is the mind ? Strictly speaking, the mind is nowhere, since it is a set, hence a conceptual object; only brains, whether minding or not, are somewhere (MB p. 207). Further, according to this model there can be no real mind-body dualism, thus avoiding an old stumbling block of philosophy, neither mind-body interaction – as opposed to brain-body interaction. MB explain in following passage (MB 1997 p. 206):

> "There can be no mind-matter interaction because – unlike individual mental processes and brains – mind and matter are sets, hence conceptual objects. However, it does make sense to speak of 'mental-bodily interactions' provided this expression is taken to abbreviate 'interactions among plastic neuronal systems, on the one hand, and either committed neuronal systems or bodily systems that are not part of the Central Neuronal System on the other'. Thus, there are interactions between sensory and motor areas, between ideational neuronal systems and external receptors, between the cortical and subcortical regions of the brain, between the brain and the endocrine and immune system, and so on. Because mental events are neural events, and because the causal relation is defined for pairs of events in concrete systems (recall Sect. 1.9), we have:
>
> Corollary 6.5.: Mental events can cause nonmental events in the same body, and v.v.
>
> Consequently, disturbances of nonmental biofunctions may influence mental states and, conversely, mental events such as acts of will may influence nonmental bodily states. This is what neurochemistry, neurology, psychiatry, psychosomatic medicine, psychoneuropharmacology, education, and propaganda are all about".

[4] Of course, in this context the notion of set, a concept of formalized logic, is not the most user-friendly one, at any rate other less formal conceptualizations are also possible. But we do believe the MB-model is the most economic model that can explain so much. For the moment we just ask the reader to bear with us, and to judge later, based on the problems that can be addressed.

In a similar vein and building on the above assumptions one can then define consciousness, termed the highest of all brain functions by Bunge and Mahner (MB 1997 p. 209). First they define (we only slightly modify their phrasing):

> **DEF-MB-2**. A *conscious mental process* / choice / act (conceived as based on, governed by, a mental process) is a mental process / choice / act that is *monitored* (recorded, analyzed, controlled, or kept track of) by some other mental activity in the same brain.

Simply put: for a mental process or act to be conscious, it must be thought about by a higher part of the brain[5]. And further (MB 1997 p. 209):

> **DEF-MB-3**. The *consciousness* of an animal b is the set of all the states (or, rather, processes) of the brain of b in which b is conscious of some perception or thought in b itself.

(This is not a circular definition thanks to DEF-MB-2.) MB explain:

> "According to this convention, an animal can only be conscious of some of its own higher mental processes: not just feeling, sensing, and doing, but also thinking of what it perceives or thinks. An animal conscious of mental process x (in itself) possibly undergoes (either in parallel or in quick succession) *two* different mental processes: x – the object mental process or content of its consciousness, and thinking about x – i.e., being conscious about x." (MB 1997 p. 208-209)

Using DEF-MB-2, one can now construct a definition of the voluntary act as invoked for instance by Aristotle, Aquinas and Ayer (MB 1997 p. 210).

> **DEF-MB-4**. An animal act is *voluntary* (or intentional) IFF it is a conscious purposeful act.

'Purposeful' is left undefined, but is self-explaining; and, for that matter, it seems less essential to us since it seems that a choice or conscious act can always be associated with a purpose. And finally (MB 1997, p. 211):

[5] To us, it seems a priori a quite acceptable hypothesis, in the light of neuroscientific findings showing that the cognitive center of the brain, the prefrontal cortex, is connected to large parts of the brain and seems to function as the 'integrator' of information embedded in very many other mental processes. This is the neurobiological intuition that lies at the basis of MB's intuition that consciousness of X is related to 'thinking about X by a higher level part of the brain' – presumably the prefrontal cortex.

DEF-MB-5: An animal acts of its own *free will* IFF

i. its action is voluntary; and

ii. it has free choice of its goal(s) – i.e., is under no programmed or external compulsion to attain the chosen goal.

In essence, we find in clause (ii) the absence of constraint that is the essential condition for Ayer and others. And 'voluntary' in clause (i) and in DEF-A (iii) is now explained by DEF-MB-4: it means conscious and purposeful. Therefore, to our satisfaction, we find that Mahner, Bunge and Ayer define free will in a very similar manner, moreover a manner that can well be related to a majority of compatibilist theories on free will, as will be shown further. Let us for the moment omit the notion of purpose in voluntary (which seems less essential and can be seen to be contained in the notion of 'choice'). Then we can synthesize Ayer's and MB's models as follows.

DEF-MBA. Action A by animal b is 'free-willed' or 'free' (is made of b's own free will) IFF

i. the action A is *unconstrained* (no programmed or external compulsion), and

ii. the action A is *conscious* in that the action (linked to a mental process) is *monitored* (recorded, analyzed, controlled, or kept track of) by some other mental activity in the brain of b.

So we have specified, with MB, that no-constraint is essentially absence of constraint by external agents and by externally programmed influences. We could now define 'free will' as the *capacity* to perform free-willed actions in the sense of DEF-MBA.

It is worth mentioning that DEF-MBA is, luckily, in a long historical lineage. Notably, it comes close to Aristotle's view on free will, but it is also meaningfully linked to Kant's (see further). Aristotle conceived of free will as a capacity to make choices that are unconstrained and *not made of ignorance* (cf. excerpts in Pereboom 1997, Ch. 1). The latter idea can well be linked to clause (ii) of DEF-MBA, as we will detail in a moment. Thus, we see appearing in this model of free will the component of consciousness that we intuitively suspected from the start to be an ingredient of free will – simply by recognizing or assuming that 'real' free will cannot be attributed to lower animals and is typically a human thing.

A second important thesis we started from is that free will comes in degrees. We posited this degree-thesis essentially by recognizing that there must be a (continuous) evolution in free will from primitive humans to modern humans, and from a new-born child to a knowledgeable adult. A logical question to ask is then: which concepts in DEF-MBA admit of degree ? Constraint, but more conspicuously the notions of *analysis* and *control* in clause (ii)[6].

6. Elaboration of the MBA-model. Is there a 'measure' of the variability of free will ?

If free will comes in degrees, a natural question is: can one then define a measure of the variability of free will ? A related question that brings this matter sharply to the point is: can we define a maximal form of free will – maximal in the sense of optimal, most adequate ? Especially the second question seems complex and touchy; it is largely absent from the modern debate, but it did peak through in the views on free will of some of the ancients and Kant. We will use the second question as a heuristic even if speculative tool to try to answer the first one.

To highlight the relevance of this problem, let us look at a case mentioned in (Griffith 2013 p. 33). Suppose a young girl, Trina, lives in a closed community that taught her from her early childhood that stealing from people outside the community is praiseworthy, the thing to do. Suppose that on her first day at school Trina happily puts her worldview to practice and steals several objects from her little schoolmates. Does Trina do these acts of her free will ? As argued by Griffith, in this case many people would accept that Trina was brainwashed (to some degree), and that she is not really blameworthy because she did not know any better. Thus, many would believe that Trina has no real free will. This is also what the MBA-model says: while stealing Trina can surely be conscious of it (she may perform some more or less conscious analysis of some of her deeds, there may be a little voice in her head saying: "yes, stealing from these girls is cool"), but at the same time there is a programmed constraint acting, in the sense that she was brainwashed by parents and community. So it is clause (i) that implies that her act of stealing is not free-willed. But again, it seems clear that the notion of *degree* of free will is helpful here: one could say that Trina's behaviour is free to some degree, but not an optimal one. She is constrained by some (harsh or mild) form of brainwashing; and, especially, she monitors, analyzes, controls her deeds by a sub-

[6] In general the 'consciousness' of clause (ii) can vary; this comes close, or can at least interestingly be compared, to the account of O'Connor (2009).

optimal hypothesis or worldview (namely that stealing is praiseworthy). To make a link with what follows: she does not seem to use the optimal assumptions or 'theory' to monitor, analyze and control her deeds.

Now to our question whether an optimal form of free will could be defined. As said, this matter is not the main concern of this article; we ponder rather superficially about this question to find inspiration for our initial problem. It seems that DEF-MBA does hint to a possibility to qualify 'adequate' free will. Answering this question amounts to construing the maximal forms of analysis and control that consciousness, our brain, can perform when 'monitoring' an act, a choice, a decision. Having analyzed cases as those of Trina and many others, we believe the most synthetic concept to identify the maximal form of analysis and control is that of 'theory'. We use 'theory' here in a very broad sense, including scientific and academic theories (ethical, philosophical, sociological, political, physical,…) but more generally also belief systems, including every-day beliefs and assumptions[7]; (coherent) bodies of information; worldviews; etc. We introduce the concept of 'theory*' to define theory in this broad sense. The link with analysis and control is then not far to seek: in a sense, it seems it may be said that one always analyzes and controls a conscious act, choice or decision *with reference to, or within,* a theory*. This seems obvious when making decisions that need the input of expert 'intellectual' knowledge, but even when making a banal choice, say whether to go to the cinema or to visit a friend on an idle Thursday evening, one 'analyzes' or 'contemplates' both alternatives within certain beliefs, using certain assumptions – for instance assumptions about the satisfaction each activity will provide. (Sure, the analysis may be barely conscious and ultra-rapid in this example; but we already agreed that being-conscious-of comes in degrees.) But to identify 'maximal analysis' and 'maximal free will', we better look not at banal cases but at ethically or intellectually demanding ones. And then it seems quite clear what optimal, most adequate analysis means: namely analysis within the 'optimal theory' – the best theory that we have or have not at hand; the best-informed or most adequate assumptions on which to base our free choice or decision. Sure, in many cases it is not clear what the best theory is, but then, in many cases it is[8]. In short, according to the above analysis maximal free will is free will using the optimal beliefs,

[7] Note that the link between 'theory' and 'assumptions' is extremely intimate: any real-world theory, say a theory from physics or ethics, is based on assumptions. Very simply put: a theory is a set of assumptions and of consequences that are logically derived from the former.

[8] If one wishes, in cases when the 'optimal theory' is unknown, one may consider 'optimal theory' as an idealized, hypothetical concept, something as a hypothetical extrapolation of existing provisional theories. In natural science for instance, it is current practice to talk about 'future better theories'; in physics mathematical theorems are formulated regarding the features of these 'more optimal' future theories, such as Bell's theorem. In sum, there seems little doubt that this notion of optimal theory is operationally useful: one often can well know which theory or assumptions are better than others (for a given end); the optimal theory (in the absolute sense) corresponds to the extrapolated, hypothetical end product of this progress. Philosophy and natural science are full of this idealized concept, even if we likely will never know the ultimate theories.

assumptions, theories*, free will that is based on using the optimal theories and assumptions in our decision making. In an ethical context: the best moral theories and principles.

Now, as a first indication that we might be on the right track, notice that this conclusion rather closely fits to what Kant thought about free will in his Foundations of the Metaphysics of Morals (1786/1983). Here Kant states, notably: "a free will and a will under moral laws become one and the same" (Kant 1786/1983 BA98, cited and discussed in Walter 2001, p. 5). Thus for Kant, real free will is, in essence, *will under moral law*. In our parlance: will in accordance with (monitored, controlled by) adequate moral assumptions / theories*.

Now talking about 'optimal theories' might frighten some philosophers; remember though that we are obviously *not* looking for criteria for optimal or even valid theories; nor assuming that there always exist theories – *in the narrow sense* – for a given context. Importantly: one may not believe in the existence of 'optimal theories' while accepting the variability of free will, the point we want to make. Indeed, we were trying, initially, to identify a measure for the variability of free will – and we used the speculative reasoning above about what 'optimal' assumptions / theories could be merely as a heuristic tool for finding an answer to the initial question. We hope that the result is clear by now: we suggest that what varies in our capacity to perform actions that are free according to DEF-MBA is the '*adequacy'* of the theories* we use to monitor these actions. In other words, the assumptions and belief systems involved in monitoring a free act are more or less adequate for guiding these acts. It is in this sense that one can say that Trina (in the above example) does not seem to use the optimal assumptions or 'theory' to monitor, analyze and control her deeds.

As said, these are first considerations on this topic of the measure of the variability of free will; in the following we will not discuss and use the complex notion of optimal free will / optimal theory anymore. But we will adopt the MBA-model with the extra assumption that, in clause (ii), *the monitoring (analyzing, controlling) of the conscious brain involves theories** (in a wide sense), and *that these theories* have an adequacy that admits of a degree*.

Here we will term our enhanced MBA-model the 'CMT-model of free will' (free will as the capacity to perform actions that are Conscious-through-Monitoring-through-Theories*, in short). In the following Section our goal is to show the potential of this model to subsume other, in particular recent, theories of free will, and solve problems of these other models. We focus on compatibilist theories as ours. As an additional benefit, we will briefly suggest how our naturalized approach can interface with new research questions in natural science (Section 8).

7. The CMT-model compared to other theories.

A first famous theory to consider is Frankfurt's 'hierarchical mesh theory' of free will (1969, 1988). In a nutshell, according to this theory an action or choice A is free-willed if it is object of, if it meshes with, a 'second-order volition' or desire – a higher desire about the first-order desire to do A. Then A is really (rationally) desired, in agreement with one's real self, with one's second-order (rational) desires; in other words A flows from the 'will one wants' – a reflective capacity animals likely do not have. A priori it seems clear that this model fits well to the CMT-model, at least that it can charitably be interpreted in resonance with our model, by noting that second-order volitions are part of the general beliefs, worldviews etc. an agent uses to guide and control his or her life and actions. Action A meshes with a second-order volition in that it is consciously monitored by (assessed, analyzed etc.) with help of a worldview, a belief system, assumptions of life, in other words theories*. So there surely seems place for a partial overlap between both models. However it is well known that cases as Trina's, the brainwashed child, are a threat to Frankfurt's theory (for a recent overview and references, see Griffith 2013, Ch. 4). Trina may well act in accordance with higher volitions, really believe in what she does, and therefore be entirely free according to Frankfurt's model – a conclusion most people would disagree with. The CMT model solves this problem: Trina is brainwashed and therefore not unconstrained; *and* she monitors (assesses) her deeds through questionable, likely inadequate beliefs. In other words, one could say she has a limited form of free will. Next, there is also a well-known infinite-regress problem threatening Frankfurt's theory (why stop at second-order and not include higher-order volitions ?). This problem is absent from our model for obvious reasons.

A next interesting and influential theory is Wolf's 'Reason View' of free will (1990). Wolf's is also a mesh theory, but whereas in Frankfurt's model free will is, in a sense, a 'subjective' matter (an act is free as long as there is a mesh between the agent's choices and his personal, subjective, higher-level desires), Wolf adds that these personal desires should also have a 'connection with the world outside' – they should have an objective dimension, they should 'connect with the True and Good' (Wolf 1990). So, in order to have free will, one should have the right, objective reasons to do things, reasons that connect to the True and Good. In a sense this theory comes quite close to ours. In our model, we would say that an action should be consciously monitored, guided by a (sufficiently) adequate *theory**, ethical or other. It seems this can be understood in Wolf's parlance as expressing that the agent should have adequate

reasons for her or his act, reasons that 'connect with the True and Good', in the sense that they are involving, embedded in, backed-up by, adequate theories*. In view of connecting our model with natural science, we believe however that 'adequate theory' is a more instrumental and precise concept than 'adequate reasons', also to interface with (computer) science, as explained briefly in Section 8 (see notably the discussion on how computer science could at least partly emulate consciousness); and we avoid the somewhat metaphorical 'connection with the True and Good'. Furthermore, it has been objected to Wolf's theory that agents can be manipulated or programmed into accessing the True and Good (e.g. Griffith 2013, Ch. 4). We avoid this problem since we assume that a free-willed act should also be free from programmed or external compulsion, via clause (i) in DEF-MBA. But as said, in spirit there is an obvious connection between Wolf's theory and ours. It may be that relying on 'theories' and 'consciousness' as we do, rather than on 'reasons' (Wolf) or 'intelligibility / rationality' as for instance Walter does (2001), is in last analysis a matter of taste (and logical construction of the theory). Still, we submit that the variability of free will, our essential starting assumption, is conceptualized most precisely via the concept of theory or rather theory*.

One of the most debated and complex recent theories on free will is Fischer's and Ravizza's 'Reasons-Responsiveness View', putting an emphasis on the 'guidance-control' involved in a free act (Fischer et al. 2007, Fischer and Ravizza 1998). We will here summarize the essence of this elaborated theory in a sketchy matter, and reserve more detailed comparison for further work. In a nutshell, an agent has free will if her actions and choices are sensitive to, respond to, reasons, where it is emphasized that this reasons-responsiveness should not rely on luck, once in a while, and that it is *not* responsiveness under compulsion or neurotic disorder. Rather, the agent should be responsive to reasons through a systematic (cognitive) mechanism that ensures guidance-control over her actions. To make the link with the CMT-model, first note that the concept of conscious control is explicitly mentioned in our basic starting model, the MBA-definition (clause (ii)). The conscious monitoring (analyzing, controlling) of an act via adequate beliefs and theories thus seems to imply reasons-responsiveness and guidance-control (or it could be construed to imply this). So there is a partial overlap of the theories. However, it seems that Fischer's and Ravizza's Reasons-Responsiveness View is subject to the same criticism as Frankfurt's theory: it seems to imply that the brainwashed girl Trina steals of her free will, since she may well act from a practical reasoning mechanism. This is a worry for this model we are immune against (cf. above). However, the comparison between the theories can be made in much more detail, as will plan to show elsewhere.

Finally, Walter (2001) has proposed a 3-component theory of libertarian free will, based on an extensive literature review, and with the aim to compare the ingredients of the model to the neuroscientific investigations by his and other teams. Again we can only sketchily compare our minimal (2-component) model with this much more elaborate theory, but in view of the similar goals this theory is a relevant reference. A summary is given in Table 1.1., p. 43 (Walter 2001). Each main ingredient (briefly, alternativism; intelligibility / rationality; and agency / origination) can be understood according to Walter's classification in a minimal, moderate or maximal version. Comparison shows in a straightforward manner that our CMT-model corresponds to Walter's alternativism in its minimal interpretation, plus intelligibility in its moderate interpretation. However there is a difference: our model does not explicitly include agency. But we define the free act of an *agent*, and therefore agency can be understood as implicit in this notion of agent *and*, especially, as made more explicit by our 'consciousness' clause (ii). Genuine agency would then correspond to the capacity of a subject to act freely in the sense we define. Clearly, there may still be advantages to make agency explicit, in order to investigate certain questions (although it may be that this move is more relevant for a libertarian stance). But as said, we wish to present here a minimal model.

8. Conclusion.

Before resuming the results presented in this article, let us succinctly indicate some avenues of research suggested by the CMT-model in neuroscience and computer science. Of course, within the naturalized tradition, we consider it a merit of a philosophical model if it can interface with natural science. We conjecture that our model could be instrumental in tackling questions related to consciousness – considered an essential but at the same time highly elusive concept in neuroscience (Stern 2017). Notably, our model allows to conceptualize some aspects of consciousness and free will that could have an empirical basis. We think here in the first place of the process of monitoring by a neuronal superstructure, presumably in the prefrontal cortex, that should represent a theory*. We believe it would be interesting to search for the neuronal correlates for 'assumptions' in primates: one conjectures that they are related to memory-circuits, or to the 'mirror neurons' that have become fashionable lately[9]. Next, it would be interesting to analyze the well-known Libet experiment through the lens of our model, as we will do

[9] It is interesting that these mirror neurons are the base for the 'theory of mind' that neurobiologists have attributed to primates as the cognitive base for recognizing the 'self' and 'the other'.

elsewhere. In IT and computer science, a much debated question is: can future computers and robots simulate consciousness and/or free will ? If possible at all, our model suggests that one of the key properties a computer should have to emulate consciousness, or to approximately mimic it, is the capacity to 'use' higher-order theories – and this notably includes the capacity to adequately apply theories to (all) real-world situations and to act accordingly. Some will conclude we are very far from this possibility. This suggests the following line of research: can machines learn to acquire and use theories*, and which types and how ? Interestingly, very recently computer scientists and cognitive scientists have indeed come to the conclusion that mastering theories is a key goal for artificial intelligence. In the words of Lake et al. (2017, abstract):

> "We review progress in cognitive science suggesting that truly human-like learning and thinking machines will have to reach beyond current engineering trends in both what they learn and how they learn it. Specifically, we argue that these machines should (1) build causal models of the world that support explanation and understanding, rather than merely solving pattern recognition problems; (2) ground learning in intuitive theories of physics and psychology to support and enrich the knowledge that is learned; and (3) harness compositionality and learning-to-learn to rapidly acquire and generalize knowledge to new tasks and situations."

And (Lake et al. 2017, p. 2):

> "The alternative approach [alternative to pattern recognition] treats models of the world as primary, where learning is the process of model-building. Cognition is about using these models to understand the world, to explain what we see, to imagine what could have happened that didn't, or what could be true that isn't, and then planning actions to make it so. The difference between pattern recognition and model-building, between prediction and explanation, is central to our view of human intelligence. Just as scientists seek to explain nature, not simply predict it, we see human thought as fundamentally a model-building activity."

Our model, then, based on well-known analytic models and on the systematic and wide-ranging work on the conceptual foundations of psychobiology by Mahner and Bunge (1997), stipulates that a free act is unconstrained (no second agent nor programmed manipulation), and consciously monitored through a theory*. It thus highlights the rational aspect of free will. Further we emphasized the fact that free will comes in degrees, a hypothesis that seems natural within a biophilosophical approach. We submitted that the concept that can most synthetically and precisely address this variability of free will is that of theory* – or rather the capacity to use theories*.

Our model is compatibilist, and appears to be independent of whether the universe is ultimately deterministic or indeterministic. If our universe would be deterministic (a hypothesis for which we

presented arguments in Vervoort 2013), here is an example of how to apply the CMT-model. According to our compatibilist position the following three facts may be true at the same time:

(i) a murder may be committed by an act of free will (in that the act was unconstrained and accompanied by a conscious reflection based on some beliefs, assumptions etc.);
(ii) but the free will of the murderer is likely to be of a corrupted type (her moral theory* is likely not adequate);
(iii) and at the same time the murder *had* to happen (in a truly deterministic universe, there are no alternatives 'on the atomic scale', taking all facts about the universe into account).

An 'ontic CMT compatibilist' – someone who assumes both determinism and that humans can have a form of free will as construed by the CMT-model – will stipulate that claims (i) and (ii) should not be adopted without *also* considering (iii); that (iii) should *seriously be taken into account*. Of course, such a compatibilist ontology may have weighty, and all but trivial, implications for our philosophy of, and our living within, society – maybe notably for our legal, punitive and educational practices. In other words, how to *apply* this type of compatibilist theory seems an extremely complex matter. But one should remember that there exist already influential schools of thought that have theorized the values and consequences of this ontology. One of the oldest and best-developed is Spinozism.

Let us also note that the hypothesis of determinism cannot only be juxtaposed to the hypothesis of free will; it can also be linked more directly to the different ingredients of the CMT-model. Notably, the theories* that accompany conscious actions are, within a deterministic worldview, acquired by determined, necessary processes; their acquisition, interpretation, application etc. may for instance be triggered or influenced by upbringing, social background, life-changing encounters – in any particular case by a potentially quasi-infinite number of particular causes.

And yet, within our model there is room for agency, notably through the capacity we have to *learn*, to improve our beliefs and our capacity to act accordingly, to adopt more adequate theories etc.. Our view is all but pessimistic or defeatist (Spinoza's theory is all but defeatist); rather, we interpret it as giving directions on how to become "freer". Namely by making efforts to acquire beliefs, views and theories that help us to deal better with this world.

Acknowledgements: We would like to thank, for expert feedback, the participants of the conference “Free Will and Causality” at the University of Dusseldorf (September 2019), in particular Maria Sekatskaya, Laura Ekstrom, Nadine Elzein, Timothy O’Connor, Alexander Gebharter, as well as Giacomo Andreoletti for careful reading of the manuscript..

References.

Ayer, A.J. (1954 / 1997). “Freedom and Necessity,” in Pereboom (1997), ed., 110-118.

Brembs, B. (2011). “Towards a scientific concept of free will as a biological trait: spontaneous actions and decision-making in invertebrates”, Proceedings of the Royal Society B 278, 930–939

Fischer, J. M., Kane, R., Pereboom, D., & Vargas, M. (2007). Four Views on Free Will. Wiley Blackwell.

Fischer, J. M. and Ravizza, M. (1998). Responsibility and Control. Cambridge: Cambridge University Press.

Frankfurt, H. (1969). “Alternate Possibilities and Moral Responsibility,” Journal of Philosophy 66, 829–39.

Frankfurt, H. (1988). The Importance of What We Care About. Cambridge: Cambridge University Press.

Griffith, M. (2013). Free Will, the Basics. London: Routledge.

Kane, R. (2005). A Contemporary Introduction to Free Will. New York: Oxford University Press.

Kane, R. (2012). The Oxford Handbook of Free Will. New York: Oxford University Press.

Kant, I. (1786/1983). Grundlegung zur Metaphysik der Sitten [Foundations of the Metaphysics of Morals], Weischedel, W. (Ed.), Complete works in 10 Vols., Vol. 6. Darmstadt: Wissenschaftliche Buchgesellschaft.

Ladyman, J., Ross, D., with contributions of Spurrett D., Collier, J. P. (2007). Every Thing Must Go: Metaphysics Naturalized. Oxford: Oxford University Press.

Lake, B., T. Ullman, J. Tenenbaum, S. Gershman, (2017). “Building machines that learn and think like people”, Behavioral and Brain Sciences, Vol. 402017, e253

Mahner, M., Bunge, M. (1997). Foundations of Biophilosophy. Berlin: Springer Verlag

Mele, A. (2009). Effective Intentions: The Power of Conscious Will. Oxford: Oxford University Press.

O'Connor, T. (2009). “Degrees of Freedom,” Philosophical Explorations 12 (2), 119-125.

O'Connor, T., Franklin, C. (2018), "Free Will", The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2018/entries/freewill/>.

Pereboom, D., ed., (1997). Free Will. Indianapolis: Hackett Publishing

Soon, C.S., Brass, M., Heinze, H.-J., Hayner, J.D. (2008). "Unconscious determinants of free decisions in the human brain", Nature Neuroscience 11, 543–545.

Stern, P. (2017). "Neuroscience: In Search of New Concepts", Science, 358 (6362), 464-465

Vervoort, L. (2013). "Bell's Theorem: Two Neglected Solutions", Foundations of Physics 43, 769-791

Walter, H. (2001). Neurophilosophy of Free Will. From Libertarian Illusions to a Concept of Natural Autonomy. Camebridge: MIT Press

Watson, G., ed., (2003). Free Will. 2nd ed. Oxford: Oxford University Press.

Widerker, D. and McKenna, M. (eds.) (2003). Moral Responsibility and Alternative Possibilities, Burlington: Ashgate Publishing Company.

Wolf, S. (1990). Freedom Within Reason. Oxford: Oxford University Press.

Wuethrich, C. (2011). "Can the world be shown to be indeterministic after all?" In Probabilities in Physics, ed. C. Beisbart and S. Hartmann, pp. 365-389. Oxford: Oxford University Press.